Audio source separation with multiple microphones on time-frequency representations

Hiroshi Sawada

NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

This paper presents various source separation methods that utilize multiple microphones. We classify them into two classes. Methods that fall into the first class apply independent component analysis (ICA) or Gaussian mixture model (GMM) to frequency bin-wise observations, and then solve the permutation problem to reconstruct separated signals. The second type of method extends non-negative matrix factorization (NMF) to a multi-microphone situation, in which NMF bases are clustered according to their spatial properties. We have a unified understanding that all methods analyze a time-frequency representation with an additional microphone axis.

Keywords: Source separation, Short-time Fourier transform, Time-frequency representation, Independent component analysis, Gaussian mixture model, Non-negative matrix factorization, Permutation problem

1. INTRODUCTION

Many research efforts have been devoted to separate audio sources mixed in a real-room situation.\textsuperscript{1, 2} If the mixed signals are captured with multiple microphones, we can exploit the spatial diversity of the sources, and therefore expect good separation.

In this paper, we explain various audio source separation methods that utilize multiple microphones. Particularly, we focus on time-frequency representations that are obtained by applying a short-time Fourier transform (STFT) to the observed signal. Then, observed signals at multiple microphones are represented as a tensor that has frequency, time, and microphone axes, as shown in Fig. 1.

The methods described in this paper fall into two categories depending on along which axis the tensor of the observed signals is sliced in order to apply a basic statistical tool. The methods in the first category (shown in the top part of Fig. 1) slice the tensor in a frequency bin-wise manner and apply independent component analysis (ICA)\textsuperscript{3, 4} or a Gaussian mixture model (GMM)\textsuperscript{5} to the sliced matrices with a time frame axis and a microphone axis. The methods in the second category (shown in the bottom part of Fig. 1) slice the tensor in a microphone-wise manner and apply non-negative matrix factorization (NMF)\textsuperscript{6, 7} to the sliced matrices with a frequency bin axis and a time frame axis.

The methods in the first category need to resolve the permutation ambiguities\textsuperscript{8} that inherent in the ICA or GMM solutions. There have been methods designed to resolve the ambiguities as a post-processing.\textsuperscript{9, 10} As more advanced approaches, independent vector analysis (IVA)\textsuperscript{11-13} and unified statistical models\textsuperscript{14, 15} have been proposed for resolving the permutation ambiguities simultaneously with ICA and GMM, respectively.

The methods in the second category need to cluster frequency sound patterns extracted as NMF bases for each source. Multichannel NMF\textsuperscript{16, 17} is a way of estimating the information related to the source location for each NMF basis, and clustering the NMF bases according to the estimated information.

This paper is organized as follows. The next section formulates the source separation problem. Sections 3 and 4 describe the two categorized methods. Section 5 concludes this paper.
2. PROBLEM FORMULATION

2.1 Mixing Process

Let $s_1, \ldots, s_N$ be $N$ source signals and $x_1, \ldots, x_M$ be $M$ microphone observations. The convolutive mixture model is formulated as

$$x_m(t) = \sum_{n=1}^{N} h_{mn} * s_n(t), \quad m = 1, \ldots, M,$$

where $t$ represents time, and $h_{mn}$ represents the impulse response from source $s_n$ to microphone $x_m$. If we apply a short-time Fourier transform (STFT) with a sufficiently long window to cover the main part of the impulse responses, the convolutive mixture model (1) can be approximated as an instantaneous mixture model:\textsuperscript{18, 19}

$$x_{ijm} = \sum_{n=1}^{N} h_{in} s_{ijn}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad m = 1, \ldots, M,$$

where $i$ and $j$ represent the frequency-bin index and time frame index, respectively. The number $I$ of frequency bins is defined according to the sampling frequency and the STFT window length. The number $J$ of time frames is defined according to the amount of STFT shift and the length of the microphone observation signals.

In this way, we obtain the time-frequency representations, Eq. (2), of the microphone observations. They can be compactly represented with an $I \times J \times M$ complex-valued tensor $X$, $[X]_{ijm} = x_{ijm}$, as the left side of Fig. 1 shows. Here we define some vector notation for conciseness: $x_{ij} = [x_{ij1}, \ldots, x_{ijM}]^T \in \mathbb{C}^M$ for $M$ microphone observations, and $h_{in} = [h_{i1n}, \ldots, h_{iMn}]^T \in \mathbb{C}^M$ for the frequency response from source $n$ to all the microphones at frequency bin $i$. Then, Eq. (2) becomes

$$x_{ij} = \sum_{n=1}^{N} h_{in} s_{ijn}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J.$$
2.2 Source Separation

The purpose of audio source separation is to obtain \( N \) separated signals \( y^{(n)}_m, n = 1, \ldots, N \) that correspond to \( N \) source signal components observed at a microphone \( m \). In this paper, the separation process is conducted for a time-frequency representation to obtain \( y_{ijmn} \), which is a separated signal component for source \( n \) at frequency bin \( i \), time frame \( j \), and microphone \( m \). By applying inverse STFT to \( y_{ijmn} \) for all \( i, j \), we obtain a time domain separated signal \( y^{(n)}_m(t) \). Let us define a vector \( y^{(n)}_{ij} = [y_{ij1n}, \ldots, y_{ijMn}]^T \in \mathbb{C}^M \) again for conciseness.

3. METHODS BASED ON FREQUENCY BIN-WISE SEPARATION

This section describes source separation methods based on frequency bin-wise operations. As shown in the top half of Fig. 1, in such methods, we obtain an \( M \times J \) matrix \( X_i = [x_{i1}, \ldots, x_{ij}] \) for each frequency bin \( i \) by slicing the observation tensor \( X \) of a time-frequency-microphone representation.

3.1 Linear Filters

First, we describe the source separation method that involves constructing linear filters in a frequency bin-wise manner. If we obtain prior knowledge about the locations of sources or the time periods when a particular source is absent, beamforming techniques can be utilized. However, if such prior knowledge is unavailable, we generally employ ICA\(^3,4\) to construct the linear filters.

ICA obtains separated signals \( \hat{y}_{ij} = [\hat{y}_{ij1}, \ldots, \hat{y}_{ijN}]^T \in \mathbb{C}^N \) by a linear transformation

\[
\hat{y}_{ij} = W_i x_{ij}, \quad j = 1, \ldots, J \tag{4}
\]

with a separation matrix \( W_i \) of size \( N \times M \). The matrix \( W_i \) is optimized so that the distribution of the vector elements \( \hat{y}_{ijn}, j = 1, \ldots, J \) is far from a Gaussian distribution. Various optimization learning rules have been proposed, for example FastICA\(^3\) or one based on natural gradient.\(^4,20\) In the learning rule, a non-linear function is utilized to evaluate how similar or different the distribution of each element \( \hat{y}_{ijn} \) is compared with the Gaussian distribution. For a complex-valued variable obtained as a result of STFT, a polar coordinate based non-linear function\(^21\) is effective.

There is scaling ambiguity in an ICA solution. For an audio source separation task, the scaling ambiguity is resolved by trying to represent the observed signals at microphones with scaled separated signals. For that purpose, we calculate the inverse matrix \( A_i = [a_{i1}, \ldots, a_{iN}] = W_i^{-1} \) or the Moore-Penrose pseudo inverse matrix \( A_i = [a_{i1}, \ldots, a_{iN}] = W_i^+ \) of the separated matrix. And by multiplying it on both sides of (4), we have

\[
A_i \hat{y}_{ij} = \sum_{n=1}^{N} a_{in} \hat{y}_{ijn} = x_{ij}. \tag{5}
\]

Then we obtain a vector \( \hat{y}^{(n)}_{ij} \) of scaled (ambiguity-resolved) separated signals as

\[
\hat{y}^{(n)}_{ij} = a_{in} \hat{y}_{ijn}. \tag{6}
\]

3.2 Time-Frequency Masking

Second, we describe a source separation method that employs time-frequency masking, in which sparseness is assumed for the frequency-domain sources \( s_{ijm} \).\(^22,23\) The sparseness of a source can be characterized by the fact that the source amplitude is rarely large and is close to zero most of the time. Frequency-domain speech or music sources are good examples of sparse sources. When such sources are mixed, we can assume that at most only one source signal makes a large contribution to each time-frequency observation. Thus, the mixture model (2) can be further approximated as

\[
x_{ijm} \approx h_{imn}s_{ijn} \tag{7}
\]

where \( n = \bar{n}(i, j) \in \{1, \ldots, N\} \) is the index of the source that has the largest amplitude, and depends on each time-frequency slot \((i, j)\).
In the method based on time-frequency masking, we need to estimate which source has the largest amplitude for each time frequency slot \((i, j)\). For that purpose, we employ a clustering method for observation vectors \(x_{ij}\) to calculate the posterior probability \(P(C_n|x_{ij})\) that a vector \(x_{ij}\) belongs to a cluster \(C_n\). Then, time frequency masks are made by

\[
M_{ijn} = \begin{cases} 
1 & \text{if } p(C_n|x_{ij}) \geq p(C_{n'}|x_{ij}), \forall n' \neq n \\
0 & \text{otherwise},
\end{cases}
\]

and separated signals are obtained by

\[
y_{ij}^{(n)} = M_{ijn}x_{ij}.
\]

In terms of clustering methods, one based on an anechoic propagation model\(^{24, 25}\) is easy and simple, and works well under low reverberant conditions. However, to cope with more complicated real-room sound propagation, frequency bin-wise clustering has been proposed.\(^{10}\) Specifically, a Gaussian mixture model (GMM)\(^{5}\)

\[
p(x_{ij}|\theta) = \sum_{n=1}^{N} \alpha_{in} p(x_{ij}|a_{in}, \sigma_{in})
\]

with a complex Gaussian density function of the form\(^{10}\)

\[
p(x|a_{in}, \sigma_{in}) = \frac{1}{(\pi \sigma_{in}^2)^{M-1}} \exp \left( -\frac{||x - (a_{in}^H x) \cdot a_{in}||^2}{\sigma_{in}^2} \right)
\]

is assumed for each frequency bin \(i\), and we estimate a parameter set \(\theta = \{a_{i1}, \sigma_{i1}, \alpha_{i1}, \ldots, a_{iN}, \sigma_{iN}, \alpha_{iN}\}\) that maximizes the likelihood \(p(X_i|\theta) = \prod_{j=1}^{J} p(x_{ij}|\theta)\) of the matrix \(X_i\). In \((10)\), \(a_{in}\) is the mean vector, \(\sigma_{in}\) is the variance, and \(\alpha_{in}\) is the mixture ratio of the \(n\)-th cluster. After the parameter set is estimated, the posterior probabilities used in \((8)\) is given by

\[
p(C_n|x_{ij}) = \frac{\alpha_{in} p(x_{ij}|a_{in}, \sigma_{in})}{\sum_{n=1}^{N} \alpha_{in} p(x_{ij}|a_{in}, \sigma_{in})}.
\]

### 3.3 Post-processing for Permutation Alignment

The method based on ICA or GMM, described in Subsections 3.1 or 3.2, performs a source separation task in a frequency bin-wise manner. Therefore, we need to align the permutation ambiguity of the ICA or GMM results in each frequency bin so that a separated signal in the time domain contains frequency components from the same source signal. This problem is well known as the permutation problem of frequency-domain BSS.\(^{8}\) Although various approaches to the permutation problem have been proposed,\(^{26}\) the following approach based on dominance measures\(^{9, 10}\) performs very well.

When using ICA to construct linear filters, we employ the power ratio

\[
r_i^{(n)}(j) = \frac{||y_{ij}^{(n)}||^2}{\sum_{n=1}^{N} ||y_{ij}^{(n)}||^2}
\]

of scaled separated signals \((6)\) as a dominance measure.\(^{9}\) On the other hand, when using a GMM for time-frequency masking, we employ the posterior probability \((12)\)

\[
r_i^{(n)}(j) = p(C_n|x_{ij})
\]

as a dominance measure.\(^{10}\) After calculating the dominance measure, we basically interchange the indices \(n\) of separated signals so that the correlation coefficient \(p(r_i^{(n)}, r_i^{(n')})\) between the dominance measures at different frequency bins \(i\) and \(i'\) is maximized for the same source. The optimization procedure is described in detail in a reference.\(^{10}\)
3.4 Simultaneous Permutation Alignment

The last subsection describes a way to solve the permutation problem by post-processing after ICA or GMM. This subsection, on the other hand, introduces some methods in which the permutation problem is solved simultaneously with the source separation procedure.

Regarding the linear filters (i.e., ICA) described in Subsection 3.1, independent vector analysis (IVA)\textsuperscript{11–13} is the corresponding method. With IVA, basically the same linear filter is constructed as with ICA (4). However, the learning rule for the separation matrix $W_i$ is different from that of ICA, and the non-linear function used in the rule depends on the separated signals of all the frequency bins $\hat{y}_{ijm}, \ldots, \hat{y}_{ijm}$.

Regarding the time-frequency masking described in Subsection 3.2, methods with unified statistical models were recently proposed.\textsuperscript{14, 15} More specifically, in addition to frequency bin-wise clustering based on a GMM, those methods employ other statistical models for the source signal activities and/or the direction of arrival of sources, and have derived inference algorithms that optimize all the model parameters simultaneously.

4. METHOD BASED ON FREQUENT SOUND PATTERN EXTRACTION

This section describes a source separation method based on frequent sound pattern extraction. More specifically, it is based on non-negative matrix factorization (NMF),\textsuperscript{6, 7} as depicted in the bottom half of Fig. 1.

4.1 Standard Single-channel NMF

This subsection reviews standard single-channel NMF. Let the observation tensor $X$ be sliced to obtain an $I \times J$ matrix $X^{(m)}$, $[X^{(m)}]_{ij} = x_{ijm}$, for each microphone $m$.

There are several distance/divergence metrics proposed for NMF. In this paper, we describe NMF with Itakura-Saito (IS) divergence, IS-NMF,\textsuperscript{27} which is recognized as being highly suited to audio modeling and separation. The NMF input should be a matrix with non-negative elements. Thus, we calculate the squared absolute value

$$\hat{x}_{ijm} = |x_{ijm}|^2 = x_{ijm}x_{ijm}^*$$

of $x_{ijm}$, and construct a non-negative matrix $\hat{X}^{(m)}$, $[\hat{X}^{(m)}]_{ijm} = \hat{x}_{ijm}$. Then, we factorize this $I \times J$ matrix into the product

$$\hat{X}^{(m)} \approx TV$$

of an $I \times K$ matrix $T$ and a $K \times J$ matrix $V$. Here, $K$ is the number of NMF bases. The matrix elements $t_{ik} = [T]_{ik}$, $v_{kj} = [V]_{kj}$ of the factored form should also be non-negative. Figure 2 shows an example of NMF applied to an audio signal. Five sound patterns are extracted as five NMF bases shown in matrix $T$. Matrix $V$ indicates when each basis becomes active with its intensity.

Algorithms for NMF minimize the divergence between the given matrix $\hat{X}^{(m)}$ and its factored form $TV$. Let us define

$$\hat{x}_{ij} = \sum_{k=1}^{K} t_{ik}v_{kj}$$

Figure 2. Frequent sound patterns are extracted by NMF as its bases.
Assign NMF bases $T$ to each output according to their spatial properties.

Figure 3. Multichannel extension of NMF achieves audio source separation by clustering NMF bases according to their spatial properties.

as the element that should be equal to $\hat{x}_{ijm}$. Then, the IS divergence between $\bar{X}^{(m)}$ and $TV$ is defined as

$$D_{IS}(\bar{X}^{(m)}, \{T, V\}) = \sum_{i=1}^{I} \sum_{j=1}^{J} d_{IS}(\hat{x}_{ijm}, \hat{x}_{ij}) ,$$  \hspace{1cm} (18)

$$d_{IS}(\hat{x}_{ijm}, \hat{x}_{ij}) = \frac{\hat{x}_{ijm}}{\hat{x}_{ij}} - \log \frac{\hat{x}_{ijm}}{\hat{x}_{ij}} - 1 .$$  \hspace{1cm} (19)

The IS divergence defined by (18) can be minimized by the following manner. First, we initialize the elements of matrices $T$ and $V$ with random non-negative numbers. Then, we apply the following update rules until convergence:

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_{j} \hat{x}_{ijm} \hat{x}_{ij}}{\sum_{j} \hat{x}_{ij}}} ,$$  \hspace{1cm} (20)

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_{i} \hat{x}_{ijm} \hat{x}_{ij}}{\sum_{i} \hat{x}_{ij}}} .$$  \hspace{1cm} (21)

These update rules are called *multiplicative*, since another non-negative value is multiplied with the element before the update.

For the multichannel extension described in the next section, here we explain a generative model related to the formulation of IS-NMF. Suppose that each of the elements $x_{ijm}$ of the observation $X^{(m)}$ is generated from a zero-mean complex Gaussian distribution $\mathcal{N}_c(x_{ijm}|0, \hat{x}_{ij})$ with variance $\hat{x}_{ij}$. Then, the maximization of the log-likelihood of the observation

$$\log p(X^{(m)}|\theta) = \sum_{i=1}^{I} \sum_{j=1}^{J} \log \mathcal{N}_c(x_{ijm}|0, \hat{x}_{ij})$$  \hspace{1cm} (22)

is equivalent to the minimization of the IS divergence defined by (18) with $\theta = \{T, V\}$.17

4.2 Multi-channel NMF (without Basis Clustering)

This subsection and the next discuss multichannel extensions of NMF. The purpose of the extension is to cluster NMF bases and to obtain source separation results as shown in Fig. 3. We begin with the simplest form where each NMF basis is attached to its own spatial property. In the next subsection, we will cluster the bases by sharing spatial properties.

For the multichannel extension of IS-NMF, we extend the generative model explained in the last subsection to a multichannel case. Suppose that each observation vector $x_{ij} = [x_{ij1}, \ldots, x_{ijM}]^T \in \mathbb{C}^M$ for $M$ microphones is...
generated from a multivariate complex Gaussian distribution $\mathcal{N}_c(x_{ij}|0, \hat{X}_{ij})$ with mean $0$ and covariance matrix $\hat{X}_{ij}$. Then, as with (22), we consider the log-likelihood
\[
\log p(X|\theta) = \sum_{i=1}^{I} \sum_{j=1}^{J} \log \mathcal{N}_c(x_{ij}|0, \hat{X}_{ij})
\] (23)
of the time-frequency-microphone tensor $X$. Now, let us consider a factored form, similar to (17), for this multichannel extension.

As Fig. 3 shows, we associate spatial properties with the NMF bases $T$. More specifically, we introduce an $M \times M$ matrix $H_{ik}$ for an element $t_{ik}$ of $T$. Then, the covariance matrix $\hat{X}_{ij}$ is modeled as
\[
\hat{X}_{ij} = \sum_{k=1}^{K} H_{ik} t_{ik} v_{kj}.
\] (24)
Matrix $H_{ik}$ should be Hermitian positive semidefinite to guarantee non-negativity in a multichannel sense. The maximization of the log-likelihood (23) is equivalent to the minimization of the multichannel IS divergence
\[
D_{IS}(\hat{X}, \{T, V, H\}) = \sum_{i=1}^{I} \sum_{j=1}^{J} d_{IS}(X_{ij}, \hat{X}_{ij}),
\] (25)
\[
d_{IS}(X_{ij}, \hat{X}_{ij}) = \text{tr}(X_{ij}^{-1}\hat{X}_{ij}^{-1}) - \log \det X_{ij}^{-1}\hat{X}_{ij}^{-1} - M
\] (26)
where $\text{tr}(B) = \sum_{m=1}^{M} b_{mm}$ is the trace of a matrix $B$, $\det$ is the determinant, and $X_{ij} = x_{ij}x_{ij}^H$ is the outer product of the observation vector. And $\hat{X}$ in (25) is a hierarchical matrix that has such outer products as its elements $[\hat{X}]_{ij} = X_{ij}$.

The divergence defined by (25) and (26) is minimized by repeating the following update rules.\(^{17}\)
\[
t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_{j} v_{kj} \text{tr}(X_{ij}^{-1}X_{ij}X_{ij}^{-1}H_{ik})}{\sum_{j} v_{kj} \text{tr}(X_{ij}^{-1}H_{ik})}}
\] (27)
\[
v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_{i} t_{ik} \text{tr}(X_{ij}^{-1}X_{ij}X_{ij}^{-1}H_{ik})}{\sum_{i} t_{ik} \text{tr}(X_{ij}^{-1}H_{ik})}}
\] (28)
These update rules reduce to those (20)-(21) of the single-channel counterpart if we assume $M = 1$, $X_{ij} = \hat{x}_{ijm}$, $H_{ik} = 1$. For updating $H_{ik}$, the algebraic Riccati equation
\[
H_{ik}^H A H_{ik} = B
\] (29)
is solved with
\[
A = \sum_{j} v_{kj} \hat{X}_{ij}^{-1}
\] (30)
\[
B = H_{ik}^H \left( \sum_{j} v_{kj} \hat{X}_{ij}^{-1}X_{ij}X_{ij}^{-1} \right) H_{ik}
\] (31)
where $H_{ik}'$ is the matrix before the update.
4.3 Multi-channel NMF (with Basis Clustering)

This subsection modifies the formulation of the multichannel NMF in order to cluster $K$ matrices $H_{i1}, \ldots, H_{iK}$ into $N$ classes. Let us introduce a new matrix $Z$ of size $N \times K$. The elements $z_{nk}$ indicate whether the $k$-th matrix belongs to the $n$-th class ($z_{nk} = 1$) or not ($z_{nk} = 0$). Replacing $H_{ik}$ in (24) with $\sum_{n=1}^{N} H_{in} z_{nk}$, we have

$$\hat{X}_{ij} = \sum_{k=1}^{K} \left( \sum_{n=1}^{N} H_{in} z_{nk} \right) t_{ik} v_{kj}.$$  

(32)

The definition of the multichannel IS divergence (26) is still valid even if we change the formulation of $\hat{X}_{ij}$ as above.

Now, to optimize the newly introduced matrix $Z$, let us redefine $z_{nk}$ as continuous values to satisfy $z_{nk} \geq 0$ and $\sum_{n=1}^{N} z_{nk} = 1$. We consider this redefinition corresponds to estimating the expectation of $z_{nk}$ according to the posterior probability $p(z_{nk} = 1|X, T, V, H)$ instead of the value $z_{nk}$ itself.

The multichannel NMF after replacing (24) with (32) is performed by the following update rules:

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_{n} z_{nk} \sum_{j} v_{kj} t_{ik} \text{tr}(\hat{X}^{-1}_{ij} X_{ij} \hat{X}^{-1}_{ij} H_{in})}{\sum_{n} z_{nk} \sum_{j} v_{kj} t_{ik} \text{tr}(\hat{X}^{-1}_{ij} H_{in})}}$$  

(33)

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_{n} z_{nk} \sum_{i} t_{ik} \text{tr}(\hat{X}^{-1}_{ij} X_{ij} \hat{X}^{-1}_{ij} H_{in})}{\sum_{n} z_{nk} \sum_{i} t_{ik} \text{tr}(\hat{X}^{-1}_{ij} H_{in})}}$$  

(34)

$$z_{nk} \leftarrow z_{nk} \sqrt{\frac{\sum_{i,j} t_{ik} v_{kj} t_{ik} \text{tr}(\hat{X}^{-1}_{ij} X_{ij} \hat{X}^{-1}_{ij} H_{in})}{\sum_{i,j} t_{ik} v_{kj} t_{ik} \text{tr}(\hat{X}^{-1}_{ij} H_{in})}}$$  

(35)

For updating $H_{in}$, the algebraic Riccati equation

$$H_{in}^H A H_{in} = B$$  

(36)

is solved with

$$A = \sum_{k} z_{nk} t_{ik} \sum_{j} v_{kj} \hat{X}^{-1}_{ij}$$  

(37)

$$B = H_{in} \left( \sum_{k} z_{nk} t_{ik} \sum_{j} v_{kj} \hat{X}^{-1}_{ij} X_{ij} \hat{X}^{-1}_{ij} \right) H_{in}.$$  

(38)

For $z_{nk}$, unit-norm normalization $z_{nk} \leftarrow z_{nk} / (\sum_{n} z_{nk})$ should follow (35) to satisfy the constraint $\sum_{n=1}^{N} z_{nk} = 1$.

Source separation results are obtained with the multichannel Wiener filter

$$y^{(n)}_{ij} = \left( \sum_{k=1}^{K} z_{nk} t_{ik} v_{kj} \right) H_{in} \hat{X}^{-1}_{ij} x_{ij}$$  

(39)

with $\hat{X}_{ij}$ being defined in (32), after $T$, $V$, $H$, $Z$ are optimized by the above update rules.

5. CONCLUSION

In this paper, we tried to provide a unified view of audio source separation with multiple microphones. Regarding experimental results related to the described methods, please refer to the corresponding references. In general, the methods in Section 3 are good at separating speech sources, whereas the methods in Section 4 are good at separating music sources. The multichannel NMF described in Subsections 4.2 and 4.3 is a more recent research result, and still requires work on, for example, reducing the computational time. We hope that many researchers are interested in audio source separation and will develop many effective methods.
REFERENCES


